Business objective

The objective is to understand the data (traffic, transaction, onsite behavior etc.) and provide 3-5 insights for the management team.

Data

The data includes randomized historical data of TourRadar's website performance from Google Analytics. It has 623076 observations with 6 variables. They include: date, path, sessions, bounces, time on page and transactions. The data dictionary is shown in table 1. There was no missing data in the provided dataset.

Table 1: Data dictionary

Variable	Variable Definition
date	date of the observation
path	URL of the visit (only tour detail pages are included)
sessions	the number of the sessions
bounces	the number of bounces
time_on_page	the average amount of time (in seconds) users spent viewing a specified page
transactions	the number of bookings

Analytics approach

The approach to analyzing the data would be through exploratory analysis, RFM modelling and time-series modelling. Exploratory analysis would be used to get descriptive statistics, high level observations and correlations. RFM modelling would be used to identify interesting links between path, transactions and date. Finally, time series modelling would be used to observe interesting trends over time.

Exploratory Analysis

The descriptive statistics is shown in table 2. It shows the date range is between June 26, 2017 to October 1, 2017. The number of sessions ranged between 1 to 587. Someone spent more than 2 hours on a page. The largest number of transactions was 108.

Table 2: Descriptive Statistics

	Date	Sessions	Bounces	Time_on_page	Transactions
Min.	20170626	1	0	0	0
1st Quartile	20170722	2	0	15.2	0

Median	20170815	4	0	61.1	0
Mean	20170811	4.199	0.2759	128.8	0.3551
3rd Quartile	20170909	6	0	148.5	1
Max.	20171001	587	250	7456.1	108

There seems to be a positive correlation between sessions and bounces. This is shown in figure 1.

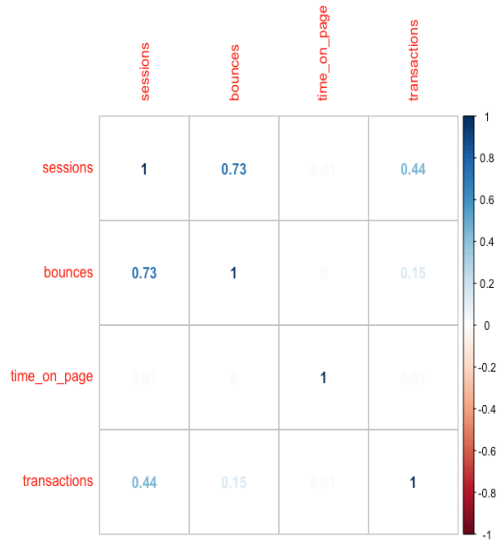

Figure 1: Correlation plot between Sessions, Bounces, Time on page and transactions

From the correlation plot we can dive deeper into the different variables. Figure 2 shows sessions with respect to bounces, time on page and transactions. Bounce rate seems to increase with the number of sessions. The number of transactions seem to be concentrations between 0 to 30 and between 1 to 300 sessions.

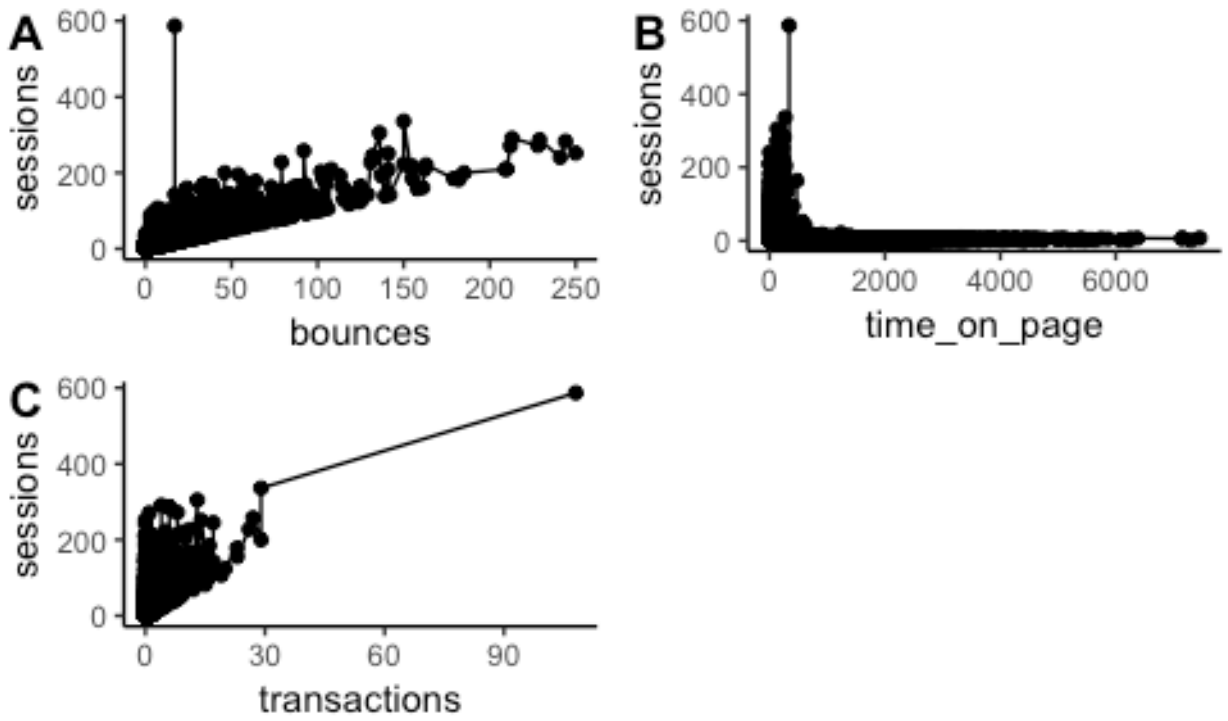

Figure 2: Sessions with respect to bounces, time on page and transactions

There is no discernable pattern for transactions and bounce rate. The bounce rate seems to increase as the time on page decreases. This is shown in figure 3.

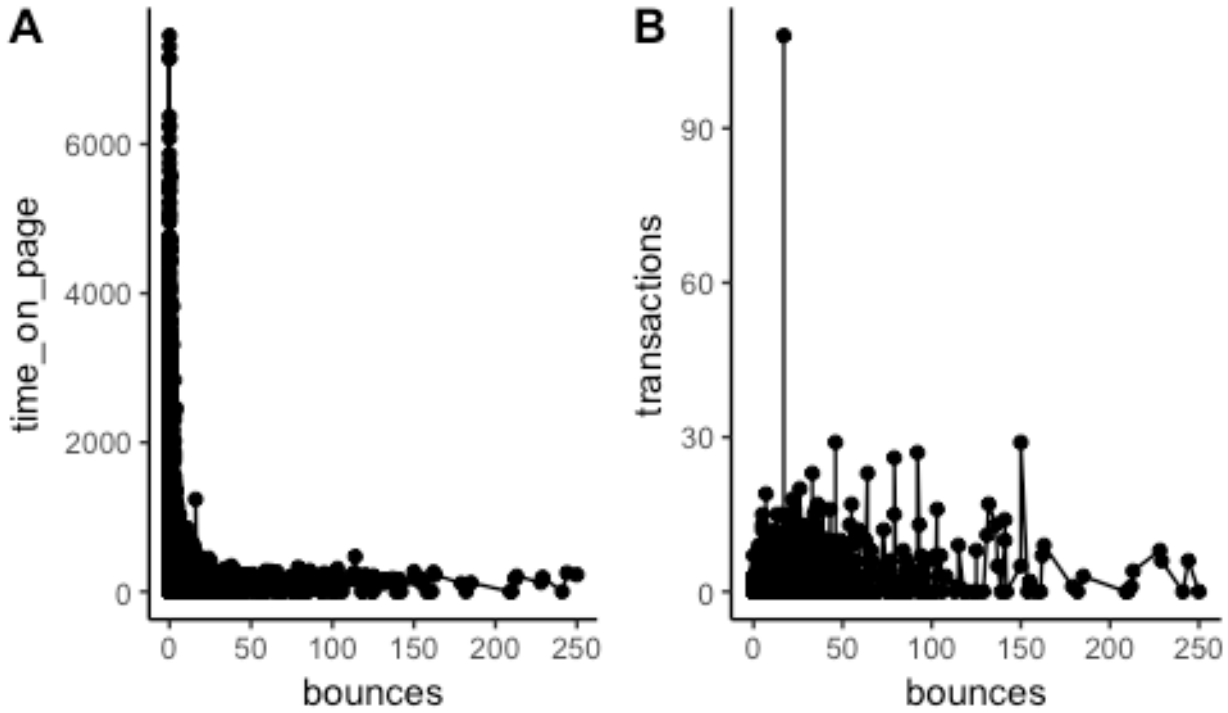

Figure 3: Bounce rate with respect to time on page and transactions

There does not seem to be a discernable pattern between time on page and number of transactions as shown in figure 4.

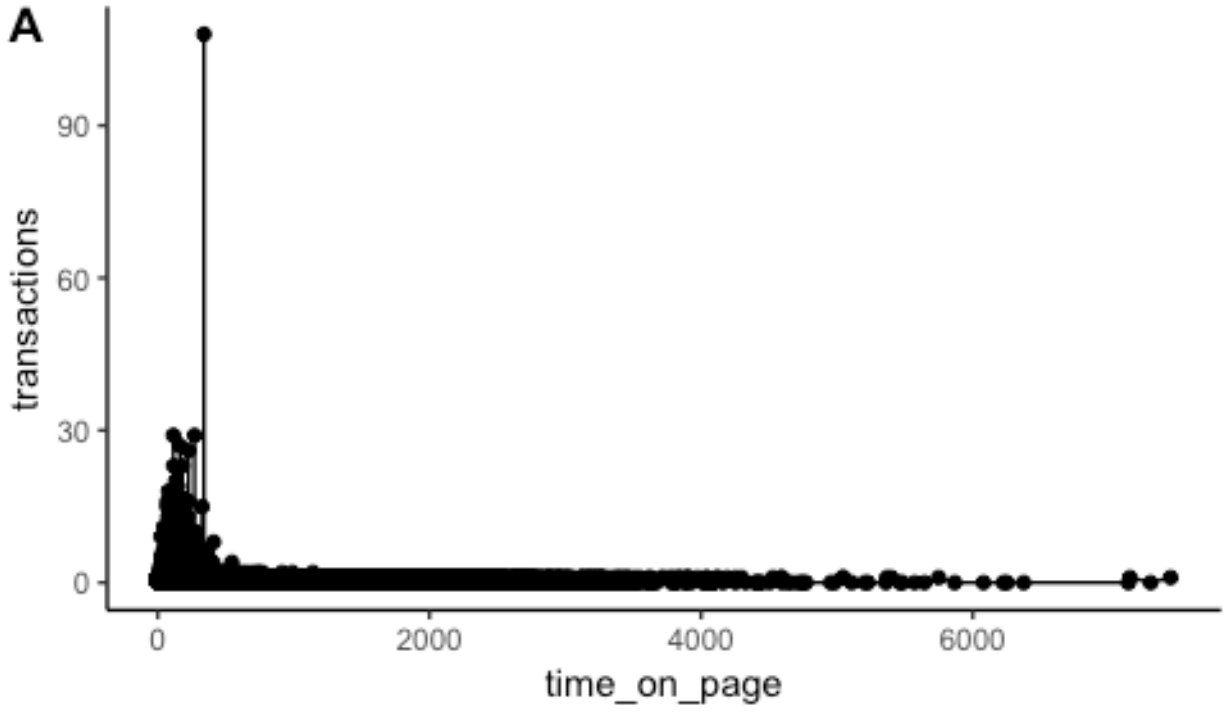

Figure 4: Time on page and transactions

RFM Modelling

The purpose of the RFM (Recency, Frequency, Monetary) model is to see the impact of the different paths. From the model, it seems very few paths bring in a high number of transactions and they are not frequently used. It is possible these are one-time bookings for a large group. A snapshot of the model is shown in Table 3.

Table 3: RFM model output

path	recency	frequency	monetary	Score
/t/14263	296	1	428	535
/t/63896	296	1	465	535
/t/59429	296	1	285	534
/t/9346	296	1	306	534
/t/74643	296	1	226	533
/t/19089	296	1	94	532
/t/14044	296	1	103	532
/t/79058	296	1	103	532
/t/91474	296	1	110	532
/t/100236	296	1	115	532

Time Series Modelling

The time series modelling is to see if there are any trends from the date perspective. This will look at the date with respect to sessions, time on page, bounces and transactions. The number of sessions seems to fluctuate mildly between July and August. There is major fluctuation in August and an irregular spike in September. This is shown in figure 5.

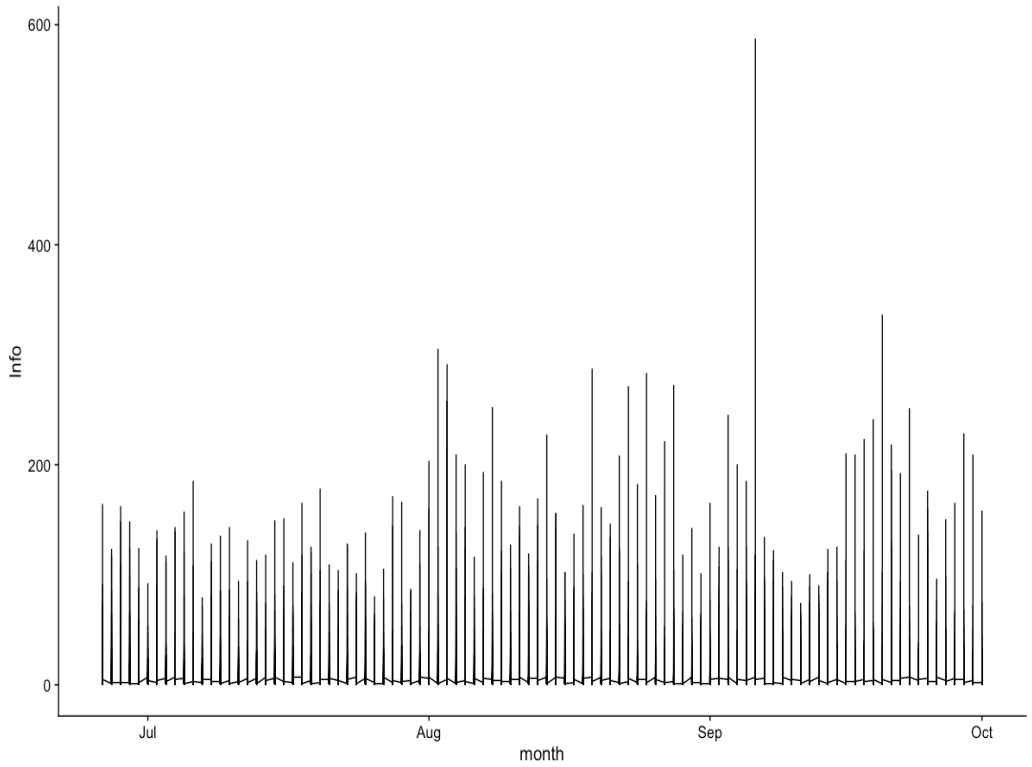

Figure 5: Sessions time series

The bounce rate had massive fluctuations between August and October which is quite alarming as shown in figure 6.

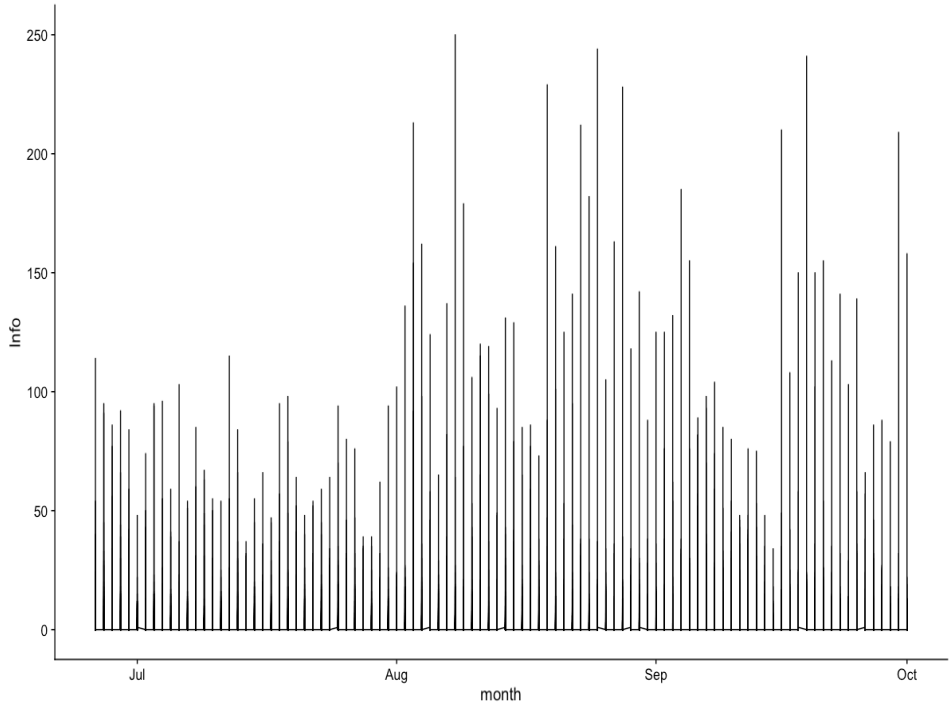

Figure 6: Bounce rate time series

The time on page seems to have the same pattern throughout the months as shown in figure 7.

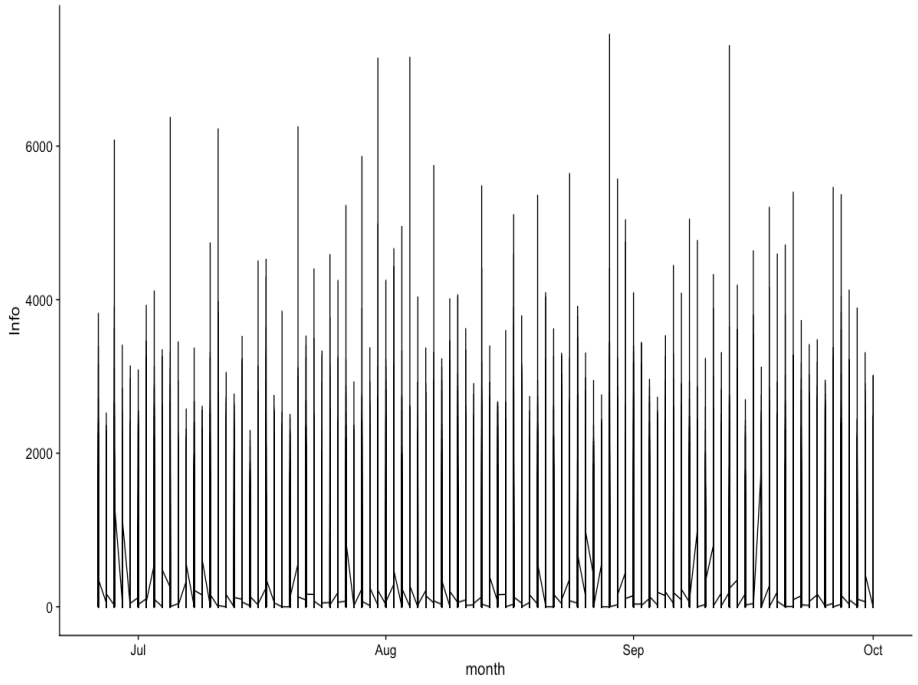

Figure 7: Time on page time series

The number of transactions seem to be consistent throughout the months but there was a very large transaction in September as shown in figure 8.

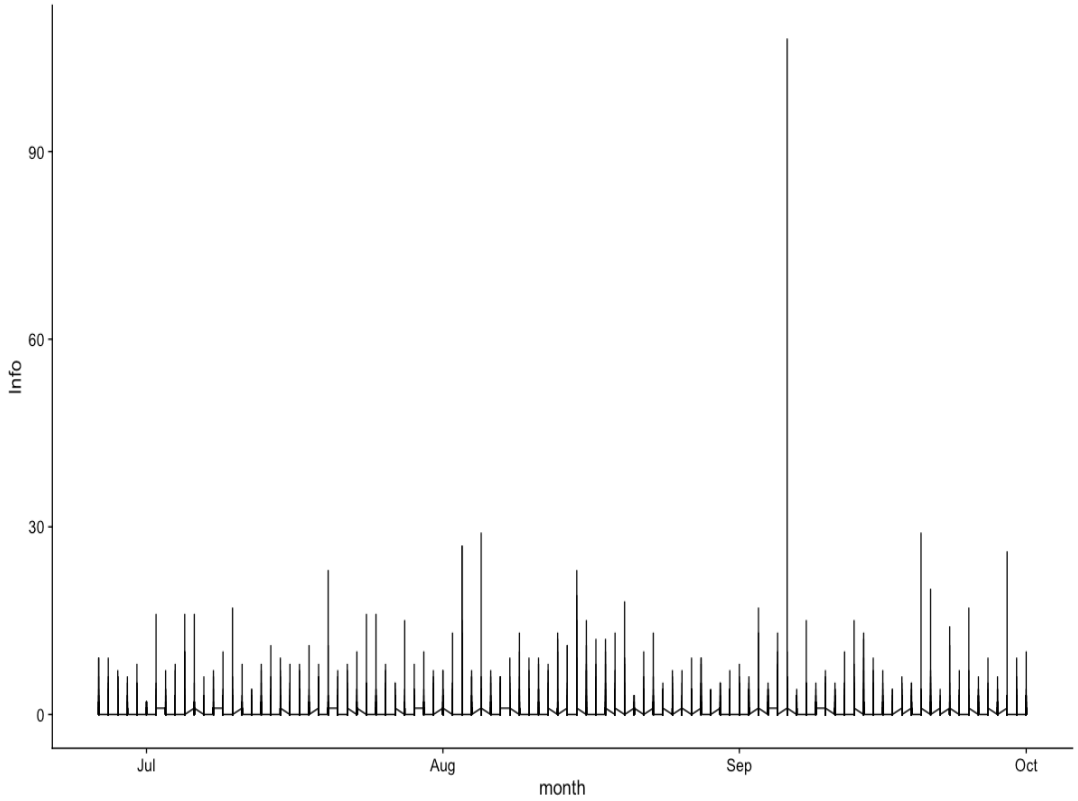

Figure 8: Transactions time series

Insights

Based on the analysis here are the key takeaways.

- Only a handful of paths generate a large number of transactions. These accounts should be prioritized
- There were a large number of bounce rates between August and October most likely driven by the increased number of sessions
- Transactions take place over a short period of time

Next steps

In terms of next steps, more granular data should be provided as Google analytics aggregates information. Furthermore, from the updated data it may be possible to design a machine learning model that would predict the number of transactions.
find out what is causing the large bounce rates between August and October. Also find out what causes the large transactions in some of the months.